Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет
по лабораторной работе №2

«Изучение библиотек обработки данных»

по курсу «Технологии машинного обучения»

ИСПОЛНИТЕЛЬ:

Голубкова С.В.
Группа РТ5-61Б

_____

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

Москва 2020

Лабораторная работа № 2

1. Цель лабораторной работы

Изучение библиотеки обработки данных Pandas.

2. Задание

Выполните первое демонстрационное задание "demo assignment" под названием "Exploratory data analysis with Pandas" со страницы курса https://mlcourse.ai/assignments

Условие задания:

https://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_english/assignments_demo/assignment01_pandas_uci_adult.ipynb?flush_cache=true

3. Выполнение работы

Лабораторная работа №2 по курсу ТМО

In [10]:
```
import numpy as np
import pandas as pd
```

In [4]:
```
data = pd.read_csv('adult.data.csv')
data.head()
```

Out[4]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

1. Сколько мужчин и женщин представлено в наборе данных?

In [15]:
```
print('{} женщин'.format(data[data['sex'] == 'Female']['sex'].count()))
print('{} мужчин'.format(data[data['sex'] == 'Male']['sex'].count()))
```

```
10771 женщин
21790 мужчин
```

2. Каков средний возраст (признак age) женщин?

In [16]:
```
print('Средний возраст женщин - {}'.format(data[data['sex'] == 'Female']['age'].mean()))
```

```
Средний возраст женщин - 36.85823043357163
```

3. Какова доля граждан Германии (признак native-country)?

In [22]:
```
data[data['native-country'] == 'Germany']['native-country'].count() / data['native-country'].count()
```

Out[22]:
```
0.004207487485028101
```

4-5. Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак salary) и тех, кто получает менее 50K в год?

In [33]:
```
print('Средний возраст богатых {}+-{} лет'.format(data[data['salary'] == '>50K']['age'].mean().astype('int64'),
                                                  data[data['salary'] == '>50K']['age'].std().astype('int64')))
print('Средний возраст бедных {}+-{} лет'.format(data[data['salary'] == '<=50K']['age'].mean().astype('int64'),
                                                 data[data['salary'] == '<=50K']['age'].std().astype('int64')))
```

```
Средний возраст богатых 44+-10 лет
Средний возраст бедных 36+-14 лет
```

6. Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)

In [40]:
```
data.loc[data['salary']=='>50K', 'education'].unique() #Не правда
```

Out[40]:
```
array(['HS-grad', 'Masters', 'Bachelors', 'Some-college', 'Assoc-voc',
       'Doctorate', 'Prof-school', 'Assoc-acdm', '7th-8th', '12th',
       '10th', '11th', '9th', '5th-6th', '1st-4th'], dtype=object)
```

7. Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.

In [54]:
```
data.groupby(['sex','race']).describe()['age']
# Максимальный возраст мужчин расы Amer-Indian-Eskimo 82 года
```

Out[54]:

sex	race	count	mean	std	min	25%	50%	75%	max
Female	Amer-Indian-Eskimo	119.0	37.117647	13.114991	17.0	27.0	36.0	46.00	80.0
	Asian-Pac-Islander	346.0	35.089595	12.300845	17.0	25.0	33.0	43.75	75.0
	Black	1555.0	37.854019	12.637197	17.0	28.0	37.0	46.00	90.0
	Other	109.0	31.678899	11.631599	17.0	23.0	29.0	39.00	74.0
	White	8642.0	36.811618	14.329093	17.0	25.0	35.0	46.00	90.0
Male	Amer-Indian-Eskimo	192.0	37.208333	12.049563	17.0	28.0	35.0	45.00	82.0
	Asian-Pac-Islander	693.0	39.073593	12.883944	18.0	29.0	37.0	46.00	90.0
	Black	1569.0	37.682600	12.882612	17.0	27.0	36.0	46.00	90.0
	Other	162.0	34.654321	11.355531	17.0	26.0	32.0	42.00	77.0
	White	19174.0	39.652498	13.436029	17.0	29.0	38.0	49.00	90.0

8. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.

In [74]:
```
print('Женатые: {}'.format(data[(data['sex']=='Male') & (data['salary']=='>50K')
                                & (data['marital-status'].isin(['Married-civ-spouse',
                                                                'Married-spouse-absent',
                                                                'Married-AF-spouse']))].shape[0]))
print('Холостые: {}'.format(data[(data['sex']=='Male') & (data['salary']=='>50K')
                                & (data['marital-status'].isin(['Married-civ-spouse',
                                                                'Married-spouse-absent',
                                                                'Married-AF-spouse'])==False)].shape[0]))
```

```
Женатые: 5965
Холостые: 697
```

9. Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?

In [79]:
```
hour_max = data['hours-per-week'].max()
people = data[data['hours-per-week']==hour_max].shape[0]
salary_50K = data[(data['hours-per-week']==hour_max) & (data['salary']=='>50K')].shape[0]
print('Максимальное количество часов работы: {}'.format(hour_max))
print('Столько часов работают: {} человек'.format(people))
print('Процент зарабатывающих много среди этих людей: {}'.format(salary_50K/people * 100))
```

```
Максимальное количество часов работы: 99
Столько часов работают: 85 человек
Процент зарабатывающих много среди этих людей: 29.411764705882355
```

10. Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).

In [89]:
```
pd.crosstab(data['native-country'], data['salary'],
            values=data['hours-per-week'], aggfunc=np.mean)
```

Out[89]:

salary	<=50K	>50K
native-country		
?	40.164760	45.547945
Cambodia	41.416667	40.000000
Canada	37.914634	45.641026
China	37.381818	38.900000
Columbia	38.684211	50.000000
Cuba	37.985714	42.440000
Dominican-Republic	42.338235	47.000000
Ecuador	38.041667	48.750000
El-Salvador	36.030928	45.000000
England	40.483333	44.533333
France	41.058824	50.750000
Germany	39.139785	44.977273
Greece	41.809524	50.625000
Guatemala	39.360656	36.666667
Haiti	36.325000	42.750000
Holand-Netherlands	40.000000	NaN
Honduras	34.333333	60.000000
Hong	39.142857	45.000000
Hungary	31.300000	50.000000
India	38.233333	46.475000
Iran	41.440000	47.500000
Ireland	40.947368	48.000000
Italy	39.625000	45.400000
Jamaica	38.239437	41.100000
Japan	41.000000	47.958333
Laos	40.375000	40.000000
Mexico	40.003279	46.575758
Nicaragua	36.093750	37.500000
Outlying-US(Guam-USVI-etc)	41.857143	NaN
Peru	35.068966	40.000000
Philippines	38.065693	43.032787
Poland	38.166667	39.000000
Portugal	41.939394	41.500000
Puerto-Rico	38.470588	39.416667
Scotland	39.444444	46.666667
South	40.156250	51.437500
Taiwan	33.774194	46.800000
Thailand	42.866667	58.333333
Trinadad&Tobago	37.058824	40.000000
United-States	38.799127	45.505369
Vietnam	37.193548	39.200000
Yugoslavia	41.600000	49.500000